A Theory of Measurement

Norman M. Bradburn (NORC and University of Chicago), Nancy Cartwright (Durham University and UCSD) and Jonathan Fuller (University of Toronto)

CHESS Working Paper No. 2016-07
[Produced as part of the Knowledge for Use (K4U) Research Project]
Durham University
September 2016

The K4U project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 667526 K4U) The above content reflects only the author's view and that the ERC is not responsible for any use that may be made of the information it contains

A Theory of Measurement

Norman M. Bradburn, Nancy L. Cartwright and Jonathan Fuller

Norman M. Bradburn
Department of Psychology
University of Chicago
1155 E. 60th St., Chicago, IL 60637
Email: bradburn-norman <at> norc.org

Nancy Cartwright
Philosophy Department
Durham University
50 Old Elvet
Durham DH1 3HN
Email: nancy.cartwright <at> durham.ac.uk

Jonathan Fuller
MD/PhD Program
University of Toronto
27 King's College Circle
Toronto, Ontario M5S 1A1
Email: jonathan.fuller <at> mail.utoronto.ca

1. What Does Mean to Measure?

This paper discusses basic issues about the nature of measurement for concepts in the social sciences and medicine, introducing a three-stage theory of measurement. In science and policy investigations we study quantities and qualities (or quality/quantity concepts)[1] and their relations in order to understand and predict the behavior of individuals/tokens displaying those quantities or falling under those concepts. What does it mean to measure a quantity (e.g. body size) or to assign a concept or category (e.g. 'underweight') to a token? In medicine, as throughout natural and social science, measurement is not just assigning categories or numbers; it is assigning values in a systematic and grounded way. This involves applying some well-grounded metric representing the quantity (e.g. body mass index (BMI)) to the token. This requires that:

1. We define the concept or quantity, identifying its boundaries, fixing which features belong to it and which do not (*characterization*).
2. We define a metrical system that appropriately represents the quantity or concept (*representation*).
3. We formulate rules for applying the metrical system to tokens to produce the measurement results (*procedures*).

The reasons we undertake a measurement project – what we want to use the measurement results for – may affect one or more of these steps. Although 1-3 are listed as separate steps to help analyze measurement processes, what happens in each stage should influence other stages. We may, for example, come to re-characterize a category on the basis of results derived relative to a candidate metrical representation of it. This may be the case with characterizations of *quality of life*, as discussed below. Or we may pick a metric system because the procedural rules for applying it are well-defined, or users know these methods better, or they are easier to implement.

All three steps - characterization, representation and procedures - need explication. For an adequate measurement, these three must line up properly together: the representation of the quantity or quality measured must be appropriate to the central features taken to characterize it; equally, the procedures adopted to carry out the measurement must be appropriate to the formal representation

[1] For those who are anti-realist about qualities and quantities in the world, measurement theory can be seen as the study of how to assign quality and quantity concepts to systems in a systematic way that allows us to engage in further practices such as prediction, comparison, and explanation. (See brief discussion below.) We shall speak here indifferently about measuring a concept or measuring the quality or quantity the concept is supposed to name.

adopted, and we should have good reason to expect that within acceptable bounds of accuracy the values assigned indicate the values we aim towards. 'Validating' a measure requires showing that it satisfies these requirements. This can be done more or less formally, more or less explicitly, and more or less successfully. But these steps should not be neglected.

Broadly speaking, there are two metaphysical attitudes towards measurement concepts: realism and nominalism. These attitudes are analogous to (and overlapping with) realist and antirealist positions, respectively regarding theoretical entities in the philosophy of science. Does a theoretical term like 'boson' or 'prion' or a concept like 'hospital quality' or 'mild cognitive impairment' refer to real features, or is the concept a *mere* name, a word used in an orderly way in language but not referring to anything in the world? If a measurement concept refers to real features, then 'representing the concept' (using a table of indicators for 'hospital quality' or a particular cognitive assessment scale for 'mild cognitive impairment') can be taken to mean representing the features to which the concept refers. If the concept does *not* refer to any real features, then representing the concept might mean something like 'representing its internal structure'. Finally, if a concept and its representation correspond to real world features, then measurement procedures seek to assign values to tokens corresponding to the level or magnitude of those features that the token displays. But if the concept and its representation do *not* refer to real features, measurement procedures my simply seek to output values that can be used in inference and that satisfy certain desiderata, like predictive success or reliable ordering of members of a set useful for other purposes. We remain neutral as to whether or not the particular measurement concepts we will discuss here are real. The three-step account of measurement we present is consistent with both realist and nominalist perspectives.

2. Three Steps

2.1 Characterizing the Concept

The characterization of concepts in the sciences may be more or less precise. The more precise the characterization, the more likely the concept is to be defined in terms of its measurement procedures (operationalization). But most concepts in social science, public health, and medicine - particularly in the policy realm - are loosely defined at the start. Further, the definitions may vary depending on the use to which the concept is put. For example, 'disability' may mean different things depending on whether we are talking about a particular individual, about a policy goal, a variable in a psychological theory, or a characteristic of a group of individuals. Concepts used in political discussions are often used in loose senses.

Many social science and medical concepts seem to refer to specific qualitative features (e.g. sex) or quantitative features (e.g. age) that individuals or populations might have, or to defined sets or functions of these features (e.g. Medicare recipients). We call these definitions *pinpoint* concepts. Other concepts sort things into categories based on a set of criteria that are loose or hard to articulate precisely, where the members of the same category need not all share any defined set of features but rather have what Wittgenstein called 'family resemblance'. For reasons explained below, we call these family-resemblance-like concepts *Ballung* concepts. 'The number of people with 'diabetes mellitus', 'mortality rate', and 'the proportion of the population with impaired hearing' seem to be pinpoint concepts. Concepts with potential normative implications and those that have evolved from everyday concepts that serve a variety of different purposes, such as 'health', 'disease', 'human welfare', 'human rights', and 'poverty', tend to be Ballung concepts. An abundance of scholarship in the philosophy of medicine has attempted to define precisely a pinpoint concept of disease (Lemoine 2013), with several fairly successful yet distinct candidates surviving, and no consensus in sight. One plausible explanation is that our everyday concept of disease is a Ballung concept that permits multiple rational reconstructions. Several diagnostic categories also seem to be Ballung concepts; e.g. 'flu-like illness', which denotes a family of infections resembling the flu.

The distinction here is not one between natural features on the one hand and ones that are socially constructed or dependent on social relations on the other hand. 'Being a stepmother' depends on what social relations obtain but is a specific, unambiguous feature. 'Race' is a paradigm for a concept generally taken to be socially constructed but that has been, in many societies, meticulously defined (e.g. based on region of origin) to allow labels to be applied unambiguously so that all the members of the same category do resemble each other in precisely the ways laid out in the definition (even though these may not have any other importance). Nor is the distinction we are trying to make that between the observable/unobservable. It would also be misleading to cast the distinction in terms of realism versus nominalism since family resemblance can certainly be based in reality, and it might be a perfectly objective fact that individuals clump into categories according to these resemblances.

Otto Neurath[2] maintained that most concepts used in daily life are of the second type. He called them *Ballungen* ('congestions'),[3] as in the German 'Ballungsgebiet' for a congested urban

[2] Neurath was a socialist, sociologist, philosopher, one of the founding members of the Vienna Circle, and spearhead of the unity of science movement of the 1930's.

[3] See Cartwright et al. (1996) for discussion and references.

area with ill-defined edges. There is a lot packed into the concept, there is often no central core without which one does not merit the label, different clusterings of features among the congestion (*Ballung)* can matter for different uses, and whether a feature counts as being inside or outside the concept - and how far outside - is context- and use-dependent. We employ Neurath's word because other words more commonly in use throughout philosophy and the sciences, like 'umbrella concept' or 'family-resemblance concept', have different meanings for different scholars in different fields.

Neurath's doctrines about Ballung concepts were influenced by Max Weber (1949). Weber argued that the study of society could probably not become a proper science because the hallmark of proper science is the use of precise, unambiguous concepts that figure in exact relations with one another. Physics, he believed, can pick and choose the concepts it studies in an effort to find such concepts but the study of society has no such latitude since it is supposed to help us understand and manage the concepts we are concerned with in the conduct of life, few of which have the right character to participate in exact science. Concepts familiar to general society, such as 'disability', 'poverty', or 'functional literacy', are bound to have multifaceted meanings, so offering a single precise characterization is likely to sacrifice or alter aspects of their meaning. This may well be the case with respect to a number of medical concepts, like 'depression', 'obesity', and 'health'.

It is essential that our measurement procedures measure the concepts we are aiming to measure, so the importance of definition cannot be overemphasized. Explicit definition is the most straightforward way to go and has become increasingly common in medicine, where explicit criteria for clinical categories are routinely decided at consensus conferences. When there is a well-articulated body of knowledge already accepted, implicit definition via the role the concept plays with respect to other concepts in a system of claims or axioms is the next tightest way to characterize a concept. This is the category that Northrop (1947) calls 'concepts of postulation'. Generally, we are not in a position to do this in the social sciences or medicine, in part because we generally lack mathematical theories/models (though it is often not very easy in the natural sciences either). After all, part of the point of measuring a concept is to find out how it relates to other concepts. Usually we need to start with some rough defeasible characteristics of the concept and through a gradual back-and-forth process refine the characterization simultaneously with refining our procedures for measuring it and our claims about its relations to other concepts.[4] The fact that we often start with rough, open-ended characterizations does not imply that the concept in view is a Ballung concept since this is the historical trajectory of many natural science pinpoint concepts like 'temperature'.

[4] For an example of this back-and-forth process from the natural sciences, see Chang (2004).

When our understanding of a concept is weak and our knowledge of what other features might serve as good indicators of it is weak as well, we sometimes resort to one kind of explicit definition: operational definition. We point to a set of relatively well-articulated measurement procedures and define the concept in terms of them. The concept is then whatever it is that these procedures assign values to. The intelligence quotient (IQ) is the canonical example: "IQ", some maintain, "is just what IQ tests measure". Another example might be body mass index (BMI). We often speak of BMI as if it is a property of individuals ("the patient's BMI is 25"). But we cannot say much more about it than that it is the mass divided by the height squared.

Operationalization makes knowledge accumulation difficult. It becomes hard to justify that other procedures measure the same quantity since that requires defending the empirical hypothesis that the new procedures yield the same values as those that define the concept. We also often gain confidence in our measurement results and in our characterization of the quantity by noting that different procedures for measuring it yield roughly equivalent results, which is difficult when quantities are defined operationally.

A good example of the difficulties of operational definition in medicine is provided by *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*), which is the premiere manual for diagnosing psychiatric disorders, particularly in North America. It operationalizes the diagnosis of mental disorders by providing lists of criteria (the presence of certain symptoms and behaviors, the absence of other disorders) that define the diagnosis in question. Operationalization supposedly increases inter-rater reliability of psychiatric diagnosis and aids in communication among healthcare professionals. But DSM categories are less helpful in research settings, presumably because there is much causal heterogeneity underlying DSM disorders. Part of the reason for this heterogeneity is that symptoms and behaviors are variably caused and multiply realized. Another explanation for the underlying heterogeneity is that many DSM diagnoses are polythetic (i.e. a patient must satisfy a certain number of criteria, none of which are individually necessary). The way that mental disorders are operationalized by the DSM creates challenges for research and the accumulation of knowledge. It becomes difficult to create scientific models of particular mental disorders and to design treatments that have large effect sizes in DSM-defined categories of patients.

No matter which of the two types[5] of concept – pinpoint or Ballung - we consider, in characterizing the concept we are usually pulled in two different directions: generality and fitness-for-purpose. Making concepts more precise can make them more fit-for-purpose but proliferates concepts and measures. Moreover, as with concepts defined operationally, purpose-built concepts make the accumulation of knowledge difficult, so we are often reasonably pulled to rely on more general concepts that have poorer fit and hope that results established in one situation are relevant for other situations.

2.2 Representation

2.2.1 Systems of Representation

Representation of pinpoint concepts is usually done using some metrical system with an underlying mathematical structure. Stevens (1951) enumerated four kinds of representations. The representation may simply be a record of the number of tokens of a concept, e.g. proportion of a population that is male or has a particular disease (often called *nominal* measures[6]); the representation may order tokens, e.g. rank hospitals on their reputation (*ordinal* measures); it may order tokens on a scale with equal intervals, e.g. blood glucose (*interval* measures); or it may order tokens on a scale with equal ratios and a true zero point, e.g. perceptual scales of loudness (*ratio* measures).

Although the distinction between pinpoint and Ballung concepts is not sharp, it is a useful distinction to keep in mind in thinking about representation. For instance, an interval or ratio scale would be inappropriate for a Ballung concept. For Ballung concepts, there are three common strategies for representation. One strategy is to shed much of the original meaning and zero in on some more precisely definable feature from the congestion that constitutes the concept. A second strategy is to represent the concept with a table or vector of features laying out the dimensions along which the family resemblances in question lie. A third strategy aims to compromise between the advantages and disadvantages of these two previous strategies by starting with a set of different indicators but then amalgamating them into a single index number.

[5] There is a third type of concept important in the sciences that we will not discuss, what might be called 'concepts of pure understanding'. These are useful for understanding, for representation (often mathematical representation), or for organization, but not for literal description.

[6] Not to be confused with taking a *nominalist* metaphysical stance towards them.

An example of the first strategy is provided by Sophia Efstathiou's (2012) discussion of the controversial Ballung concept 'race' and its introduction into different medical contexts. Efstathiou (2012) points out that the kind of variation that 'race' picks out for epidemiologists and the kind it picks out for geneticists is different in type. Epidemiologists mostly care for variation in common health outcomes and risk factors for particular diseases; 'race' in epidemiologists' speak is defined in terms of regional heritages which may then be associated with disease risks. Geneticists care about markers in the genomes of individuals and they look for interesting patterns at the level of molecular variation; their characterization of 'race' maps onto sets of genetic polymorphisms. This is a case where a loaded Ballung concept, 'race', is used in the service of different scientific research projects. Most of its content is discarded in each context, and a narrower and more precise characterization is fitted to the scientific questions being asked. This leads to a proliferation of concepts, each fit for a different discipline and purpose. Different concepts may share the same name but are not equivalent, so results established in one setting cannot be transferred to the others.

Medical 'quality of life' (QOL) is a Ballung concept *par excellence*. It is made up of many domains, each with their own characterization, with no specific delineation of its borders. The concept has thus spawned a proliferation of measures. While each individual domain, such as depression, anxiety, pain, mobility, etc., may be well measured in principle (that is, its characterization, representation and procedures are well worked out), differing domains may be included in the concept depending on the purpose to which the QOL measure is put. A treatment for a certain neurological disease might aim to improve mobility or cognition, while a treatment for pain disorder will aim to alleviate pain. The goal of improved QOL may be common to the different studies, but different measures may be used for QOL depending on the aims of the treatments (e.g. improved mobility vs. reduced pain). Consequently, it would be impossible to compare QOLs across different kinds of treatments unless it were clear that the domains underlying the QOL indices were the same.

The second strategy is to represent Ballung concepts with a table or vector of indicators, illustrated by ratings of hospital quality. A well-known measure of hospital quality (Hill & Winfrey 1996) has 3 tiers: *structure, process, and outcomes*. *Structure* is represented by such things as 'ratio of interns and residents to beds', 'ratio of registered nurses to beds', 'ratio of board-certified doctors to beds', and an index of available technology. For some specialties, there is an added measure of number of procedures performed in a year (volume), and, for others, the availability of special services such as discharge or planning services. *Process*, which is difficult to measure directly, is measured by reputation, determined by a survey of specialized physicians as a proxy to quality of care. O*utcomes* are measured by such things as risk-adjusted mortality rates, readmissions, and

infection rates. The actual measures used have evolved since the beginnings of the ratings in 1993 but retain their essential tiered structure.

This strategy comes with drawbacks. Tables and vectors do not make for easy comparison, either across time or across groups. For example, *Healthy People, 2020* (http://www/healthypeople.gov) has 26 Leading Health Indicators for population health, increased from 10 indicators in *Healthy People, 2010*, with progress to be tracked over a decade. Comparison across populations is not impossible, though. High rankings on all indicators orders a group above one with low rankings on all indicators; and there may be further reasonable ways to rank groups where differences on most indicators are large or where enough indicators differ in the same direction. Nevertheless, generally speaking at best we can expect only a partial ordering. That is not a problem with the measure. It is often, as Weber urged, the problem with the concept in which we are interested. There simply is no fact of the matter about which, among a large group of European countries with mixed results on different indicators, has a healthier population.

Sometimes we try to "solve" the problem by collapsing the indicators into a single index, generally by defining some weighting scheme to produce a single outcome. The Apgar score is an example; it is an immediate postnatal screening tool to help determine how well a newborn tolerated delivery and how well she/he is faring outside the womb. The Apgar score ranges from 0 to 10 and is calculated by adding sub-scores from five equally-rated assessments: breathing effort, heart rate, muscle tone, reflexes, and skin color. As a further example, the Montreal Cognitive Assessment (MoCA) tool screens for mild cognitive impairment by testing cognitive functioning in several domains: visuospatial/executive, naming, memory/delayed recall, attention, language, abstraction, and orientation. Again, sub-scores in each domain are added to generate an overall score, but the domains in MoCA are not equally weighted.

Aggregating indicators into a single index has obvious advantages and disadvantages. On the one hand it makes comparisons and accumulation of knowledge easier. On the other, the choice of weighting scheme is often underdetermined and sometimes downright arbitrary, which opens the possibility of cherry picking just the right weightings to get some desired result, e.g. in a clinical research study with lucrative implications. In egregious cases, some of the indicators included in the composite measure might not be relevant to the outcome we care about. For instance, Goldacre (2012) describes the influential UKPDS trial that analyzed the effect of blood sugar control on diabetic endpoints. The trial showed a 12% reduction in the composite endpoint due to intensive blood sugar management in patients with diabetes. The composite endpoint included important outcomes like sudden death, heart attack, and stroke, as well as less directly relevant outcomes like renal biomarkers. On closer analysis, most of the 12% improvement in the composite outcome was

due to a reduction in the number of patients referred for laser treatment for damage to the microvasculature of the retina rather than due to improvement in the most important cardiovascular outcomes (meanwhile, there was no significant change in the number of patients experiencing vision loss).

There are further problems beyond issues of how the weights are to be selected. When the original concept is a Ballung concept, this strategy amounts to constructing a new, more manageable concept rather than informing us about the original. Of what interest is this new concept? What purposes are served by measuring it? It may be that the new concept is useful for scientific theorizing, for prediction, or for explanation – for enterprises that rely on principles involving concepts that are precise and unambiguous. In this case it is probably most useful to treat it no longer as a Ballung concept but as a pinpoint concept, since playing a role in a network of predictive principles is one of the chief grounds on which we judge concepts to pick out specific, precise features.

Conversely, it can be misleading to represent pinpoint concepts by sets of indicators or indices. This is not to say that we may not be forced to measure a pinpoint concept in a host of indirect ways, none of which suffices to zero in on it sufficiently reliably or precisely, in which case good practice would be to report the array of results. But to represent such a concept in a theoretical structure in indirect ways risks losing the opportunity of laying out any exact relations in which it figures. Doing so blurs the line between what is vague in the world and what we are uncertain of, and blurs it in an unhelpful way.

2.2.2 Representation Theorems

In this paper we offer a general account – a theory of the nature of measurement, especially in the social sciences and medicine.[7] Sometimes the term *theory of measurement* is used more narrowly to refer to concerns about the connection between the characterization and representation, and sometimes even more narrowly to the abstract characteristics of the formal method of representation.[8] For proper measurement, these abstract characteristics of the formal representation of a concept must reflect and be warranted by the characterization of what the feature or category to be represented is. For instance, if a feature is to be represented on a scale of 1 -10, that scale should be treated as a pure ordering (an ordinal as opposed to a ratio measure) if 9 units does not equal 3

[7] For a related general account see Chang and Cartwright (2008).

[8] See Suppes (1998) for an accessible introduction.

times the amount of the quantity possessed by tokens with 3 units. So the central task in designing a good measure according to the narrow sense of measurement theory is to provide a 'narrow' *representation theorem* to show that the representation proposed has formal, abstract features appropriate to the concept, as it has been characterized.

We endorse the demand for representation theorems of this sort. But we want to underline the need to produce arguments that address the more substantive aspects of the representation. For instance, suppose our procedures dictate the use of a mercury thermometer to measure temperature. This implies that temperature is represented by the height of a column of mercury, which in turn is formally represented on an interval scale. To justify the representation, we must show that an interval scale is appropriate to the kind of thing our characterization says temperature is. But that is not enough. We must also show why readings of the height of a column of mercury can indicate temperatures in the way proposed. That will involve a lot of substantive assumptions – like the assumption that mercury expands uniformly with temperature.[9] The representation theorem makes these assumptions explicit and lays out the argument that shows that column height and temperature are indeed related as presupposed in the procedures. The measure itself can be no more warranted than the assumptions required for the proof.

Unfortunately, representation theorems are often lacking in the case of clinical research measures. Stegenga (2015) discusses the example of the Hamilton Depression (HAMD) Rating Scale. The HAMD questionnaire rates the severity of depression on a scale from 0-52 through scoring patient responses to 17 questions. One question probes the degree of suicidality, with responses rated from 0-4 (0 = suicidality absent, 4 = attempts at suicide). In contrast, there are a total of 6 points available to quantify the degree of insomnia, and four points available to quantify the amount of fidgeting. Thus, a patient who had attempted suicide might score the same with respect to these three elements of depression as a non-suicidal, somewhat fidgety patient with mild insomnia. As Stegenga observes, an antidepressant might even improve HAMD scores in a clinical trial by causing patients to sleep better and fidget less (a generic sedative could also achieve this outcome). It seems unlikely that any sound representation theorem supports the use of the HAMD scale as a representation for the concept 'severity of depression'.

One place where the need for representation theorems looms large is in the construction of index numbers by weighting different indicators. There is, as we said, a great deal of pressure to do

[9] For a fascinating study of the long effort to get characterization, representation, and procedures to mesh well in the case of temperature, we again recommend Chang (2004).

this since a total ordering of the tokens measured will then be possible, whereas with tables or vectors, usually at best only a partial ordering is possible. But in this case there should be good arguments that the weightings are appropriate to the concept to be measured and that the final representation does not imply features that the concept does not have. Often weightings are not explicitly mentioned in medical measurement, which is tantamount to tacitly applying equal weightings to indicators. In quantifying the amount of morbidity for a patient, for example, we may simply count chronic diagnoses. This scheme uses a ratio scale that counts a patient with two diagnoses as having twice the amount of morbidity as a patient with one diagnosis. Each diagnosis tacitly receives an equal weighting, even though some diseases can have a greater impact on mortality, quality of life, and other outcomes compared with other diseases.

2.3 Measurement Procedures

We most commonly think of measurement in science in terms of the procedures we carry out to assign measurement values to tokens in the world. In setting up these procedures, effort should be made to ensure that they are both *accurate* and *precise*. In common parlance, 'precision' is often conflated with 'accuracy'. Here is one way to regiment the use of the words: *accuracy* is about whether measurement results agree with the true values or locate individuals in the correct category;[10] *precision* indicates how specific a measurement result is.

Where a genuine quantity is measured, the observations are often done with an instrument that is calibrated to the metrical system that represents the quantity, such as a ruler or thermometer or by a simple counter. The observations can be transformed into various metrical systems by algorithms, such as converting feet into meters or degrees Fahrenheit into degrees Celsius. In many cases, as we noted, these instruments do not look at the quantity directly but rely on some pre-established connection between the quantity to be measured and another more directly observable quantity, as for instance, with the mercury thermometer. Similarly, we use pulse to measure heartrate, assuming that the number of arterial pulses is equal to the number of ventricular contractions, and blood pressure using a blood pressure cuff. Sometimes the more immediately observed quantity will be a cause of the targeted concept, sometimes an effect, sometimes the two are correlated for some other reason. What matters is that the two quantities be linked by reliable

[10] For nominalists, accuracy can be taken to indicate that the term has been applied in accord with all the accepted norms.

regularities. Laying out and defending these regularities is one of the central tasks in designing a measurement procedure.

This gives rise to what is sometimes called 'the problem of nomic measurement' (Chang 2004, 59). To be confident that the mercury thermometer measures temperature accurately, we must be confident that mercury expands uniformly with temperature. But to establish this empirical regularity we need an independent and accurate method of measuring temperature. This problem of justification is common to all measurement methods based on empirical laws. There are several obvious ways to circumvent this problem. First, we can determine the values of the quantity we want to measure, like temperature, by another method (this only postpones the problem, since now that other method needs to be justified). Second, we can derive the empirical law from a general theory. This is not straightforward either, since the theory relied on must be empirically justified, which can be especially difficult in the social sciences and medicine where few theories are accepted uncontroversially.

Both strategies are routinely employed in evaluating the accuracy of diagnostic tests in medicine. We often measure the accuracy of a test by comparing its performance to a 'gold standard' test that is assumed to be nearly perfectly accurate. The accuracy of a d-dimer blood test for detecting a blood clot in the lungs can be measured by comparing its performance against CT angiogram. The gold standard diagnostic test is often chosen based on theory. CT angiogram is considered the gold standard for detecting blood clots in the lungs because it visualizes clots radiographically with great resolution, an assumption that depends on both medical science and physical theory.

A mild version of operationalism can also be seen as an attempt to circumvent the regress. If empirical concepts are defined by well-specified measurement operations, observational data can be fixed without reference to theories and be made secure even while theoretical concepts and laws fluctuate and develop. This tactic brings with it all the problems we have discussed of narrowness, comparability of results from different methods, and the danger that we are no longer talking about the concept that we started out to study.

Another tactic is to look for operations that depend on relatively uncontentious empirical principles or ones that should give near-enough the same results across the range of competing empirical principles that are deemed plausible. Whether these are available or not depends on the circumstance. Herbert Feigl (1970) argued that our most basic measurement operations are grounded in middle-level regularities that seem to have a remarkable degree of stability, such as Archimedes's law of the lever and Snell's law of refraction. Again, finding these regularities seems especially problematic in the social sciences and medicine, although in physiological measurement

they often exist (e.g. measuring heartrate using arterial pulse based on the typically regular association between ventricular contraction and arterial pulsation).

For psychological concepts that are messy and in principle unobservable, Campbell and Fiske (1959) advocate a multi-trait, multi-method approach to validation. Concepts can only be accepted when they can be measured by several different methods and with different representations. This is an example of the back-and-forth process of refinement – often called 'triangulation' – among characterization, representation, and design of procedures. Much work needs to be done before a proper measure, where all three components fit appropriately, is arrived at; and a lot of substantive knowledge can be involved in the process.

Coherence along a variety of desiderata seems to provide the best solution in practice to the problem of nomic measurement in medicine as well as in the natural and social sciences. Does the quantity as measured by the proposed method behave as it is expected to? Do the results cohere with those of other reasonably defended methods? Do the empirical principles needed to support the method cohere with other reasonably justified empirical principles and theories?

Another common strategy is to provide a vector or table of results from different methods, with perhaps some attempt to describe the spread of results statistically. It is important to keep in mind, however, that in this case there is a different reason for using vectors or tables than with Ballung concepts. The difference in reason can have important consequences for how we use the information thus presented and for how we proceed to develop our science and our measurement procedures since in the first instance we suppose that there is a single correct value to be ascertained and in the latter we do not.

Just as the metric system representing a concept must match the concept, the procedures that assign values must match its representation. Mere counting of number of obese individuals in a population may be adequate if we represent 'obesity' as a dichotomous category, but it is a poor procedure for assigning a value to the amount of obesity for a measure that is sensitive to degree (e.g. BMI). Measurement systems for subjective phenomena, that is, those for which there are in principle no appeals to consensus of external observation, are particularly difficult because there is no direct way of knowing that the subjective judgments are using the same scale.[11] For subjective phenomena, measurement typically starts with observations that depend on responses from individuals to (more or less) common stimuli. The responses are then represented in some metrical system with (more or less) well-defined properties. Sometimes these observations have a one-to-one

[11] Or at least some transformations of the same scale.

relation between the response and the metric, as in the case of the 'just noticeable difference' (jnd) measures of sensation; others, by response categories labeled with vague quantifiers such as "not very often", "often", or "frequently", which are then mapped onto numeric values with only ordinal properties. Often, however, the observations are combined in some (more or less) well-specified way and put forth as a measure of a complex subjective phenomenon such as an attitude or an illness experience. For an example of the development of a measure of psychological well-being see Bradburn (1969).

The procedures used to measure a social or medical concept may end up producing a measure that does not correspond to the way the concept is meant to be understood. An arguable example is the measurement of mildly elevated blood pressure, moderately low bone density, or moderately high blood cholesterol as disease ('hypertension', 'osteoporosis', and 'hypercholesterolemia', respectively).

The borders of disease categories often change, leading to changes in rates of disease. For example, while the American Heart Association (AHA) has long considered systolic blood pressure less than 140 and diastolic 90 or less as normal for adults, in 2003 the National Heart, Lung, and Blood Institute (NHLBI) set new clinical guidelines lowering the standard normal readings to a systolic pressure equal to or less than 120 and a diastolic pressure equal to or less than 80. Thus the official statistics now report higher rates of normotension (normal blood pressure) than would be found with the NHLBI guidelines. The extent to which physicians have used the new guidelines to start treatment for patients viewed as having normotension according to the AHA guidelines is unknown. The public who are attentive to their health may be confused about whether they have normal blood pressure or not.

Even when there are good measurement properties, different operations may be used for different purposes. Consider again measures of Quality of Life used to evaluate the effectiveness of different treatments. One set of procedures with good measurement properties is the Patient Report Outcomes Information System (PROMIS) developed by the National Institutes of Health (NIH). This effort arose out of concern for the large number of items used by various researchers to measure aspects of medical QOL. Many of these measures had poor measurement properties: It was unclear what characterization they were meant to be procedures for, what their relations were to other proposed procedures for measuring the "same" or "similar" concepts, nor how they correlated with other concepts of interest. As part of a larger NIH effort to improve measurement, the PROMIS project, through elaborate review processes, has categorized concepts of interest, revised items, and submitted them to scaling procedures to produce measures that have desirable properties. The scales have been standardized on large general populations and some specialized clinical

populations. That the scales are accurate representations of the concepts was established by a large and complex series of clinical studies in which the scales were correlated with clinical and patient reported assessments. These procedures are used to establish that the scale values map onto clinically meaningful assessments in regular ways. The measures then become the criteria for evaluating changes in treatment outcomes. But as noted above, different measures may be used to characterize QOL outcomes depending on the purpose of the interventions, thus making it difficult to compare levels of QOL among different groups or over time.

3. In Sum

A good measure satisfies three requirements: 1) We have a characterization of the concept that identifies its boundaries and fixes what tokens belong to the concept and what tokens do not; 2) we have a metrical system that appropriately represents the concept as characterized; and 3) we have rules for applying the metrical system to individual tokens to produce measurement results. Only if the characterization, representation, and procedures are well specified and are shown to mesh properly has a good measure been achieved.

We distinguished between pinpoint concepts that refer to a single quantity or category that can be precisely defined, and Ballung concepts that refer to things that are loosely related but for which the boundaries of the concept are not clear. These kinds of concepts need to be treated differently not only with respect to characterization but also when it comes to representation and the design of procedures.

The use of concepts for different purposes often leads to changes in definition, representation and/or procedures that disrupt the possibility of comparison and knowledge accumulation but that often makes the measures more appropriate to the aim they are supposed to serve.

References

Bradburn, Norman. 1969. *The Structure of Psychological Well-Being.* Chicago: Aldine Publishing Co.

Campbell, D.T. and D.W. Fiske. 1959. "Convergent and discriminant validation by the multitrait-multimethod matrix." *Psychological Bulletin* 56:81-105.

Cartwright, Nancy, Jordi Cat, Lola Fleck and Thomas E. Uebel. 1996. *Otto Neurath: Philosophy Between Science and Politics*. New York: Cambridge University Press.

Chang, Hasok. 2004. *Inventing Temperature: Measurement and Scientific Progress*. New York: Oxford University Press.

Chang, Hasok and Nancy Cartwright. 2008. Measurement. In *The Routledge Companion to Philosophy of Science*, edited by Stathis Psillos and Martin Curd, 367-375. London and New York: Routledge.

Efstathiou, Sophia. 2012. "How ordinary race concepts get to be usable in biomedical science: an account of founded race concepts." *Philosophy of Science* 79 (5):701-713.

Feigl, Herbert. 1970. The 'Orthodox' View of Theories: Remarks in Defense as well as Critique. In *Analyses of Theories and Methods of Physics and Psychology*, edited by Michael Radner and Stephen Winokur, 3-16. Minneapolis: University of Minnesota Press.

Goldacre, Ben. 2012. *Bad Pharma: How Drug Companies Mislead Doctors and Harm Patients*. New York: Faber and Faber.

Hill, Craig and Krishna Winfrey. 1996. *The 1996 Index of Hospital Quality.* Chicago: NORC at the University of Chicago.

Lemoine, M. 2013. "Defining disease beyond conceptual analysis: an analysis of conceptual analysis in philosophy of medicine." *Theoretical Medicine and Bioethics* 34 (4):309-325.

Northrop, F.S.C. 1947. *The Logic of the Sciences and the Humanities.* New York: Meridian Books, Inc.

Stegenga, Jacob. 2015. "Measuring effectiveness." *Studies in History and Philosophy of Biological and Biomedical Sciences* 54:62-71.

Stevens, S. S. 1951. Mathematics, measurement, and psychophysics. In *Handbook of Experimental Psychology*, edited by S.S. Stevens. New York: Wiley.

Suppes, Patrick. 1998. Measurement, Theory of. In *The Routledge Encyclopedia of Philosophy*, edited by E. Craig. London: Routledge.

Weber, Max. 1949. Objectivity. In *The Methodology of Social Sciences*, edited by E. A. Shils and H. A. Finch. Glencoe, Ill.: Free Press.